Awake reactivation predicts memory in humans

Bernhard P. Staresina\textsuperscript{a,b,1}, Arjen Alinka\textsuperscript{a}, Nikolaus Kriegeskorte\textsuperscript{a}, and Richard N. Henson\textsuperscript{a}

\textsuperscript{a}Medical Research Council Cognition and Brain Sciences Unit, Cambridge CB2 7EF, United Kingdom; and \textsuperscript{b}Department of Psychology, Stanford University, Stanford, CA 94305

Edited by Andrew P. Yonelinas, Department of Psychology, University of California, Davis, CA, and accepted by the Editorial Board November 12, 2013 (received for review June 24, 2013)

How are new experiences transformed into memories? Recent findings have shown that activation in brain regions involved in the initial task performance reemerges during postlearning rest, suggesting that “offline activity” might be important for this transformation. It is unclear, however, whether such offline activity indeed reflects reactivation of individual learning experiences, whether the amount of event-specific reactivation is directly related to later memory performance, and what brain regions support such event-specific reactivation. Here, we used functional magnetic resonance imaging to assess whether event-specific reactivation occurs spontaneously during an active, postlearning delay period in the human brain. Applying representational similarity analysis, we found that successful recall of individual study events was predicted by the degree of their endogenous reactivation during the delay period. Within the medial temporal lobe, this reactivation was observed in the entorhinal cortex. Beyond the medial temporal lobe, event-specific reactivation was found in the retrosplenial cortex. Controlling for the levels of blood oxygen level-dependent activation and the serial position during encoding, the data suggest that offline reactivation might be a key mechanism for bolstering episodic memory beyond initial study processes. These results open a unique avenue for the systematic investigation of reactivation and consolidation of episodic memories in humans.

Episodic memory, our ability to mentally relive past events and experiences, is a fundamental property of the human mind. Although memory research in humans has largely focused on the study (encoding) and test (retrieval) components of memory paradigms, comparatively little is known about the processes occurring between those stages. Experiments recording brain activity during postlearning sleep found that the activity in regions engaged during awake task performance reemerges during sleep (1, 2). Moreover, externally reinstating a study context during sleep (presenting an olfactory stimulus that was present during encoding) has been found to bolster postsleep memory performance (3). However, whereas the beneficial role of sleep for memory is undisputed (for recent reviews, see refs. 4 and 5), it is unlikely that sleep is the sole state in which memory solidification occurs. Indeed, recent functional magnetic resonance imaging (fMRI) studies in humans showed that activation in task-specific regions can also be observed during offline wake periods and that the amount of postlearning activation in these task-specific regions correlates with later behavioral performance (6, 7).

Although these data show that activation in task-specific regions during rest periods is related to later memory performance, direct evidence for the notion that individual experiences are reactivated during these offline periods is lacking. That is, although it has been shown that experimental reactivation of individual events (by presenting sounds associated to unique study material) bolsters later memory for these events (8, 9), it is an open question whether such event-specific reactivation occurs spontaneously, and which brain regions may be involved in such reactivation. Rodent work, on the other hand, has furnished compelling evidence for endogenous reactivation of discrete study content during offline periods. For instance, the hallmark finding that hippocampal place cells that fired together during the initial experience show a tendency to coactivate again during postlearning sleep has given rise to the concept of episodic “replay” (10), a phenomenon also observed during awake rest periods in subsequent studies (11, 12). Apart from the hippocampus, cells in the lateral entorhinal cortex were recently discovered that fire in an object-specific manner at locations where these objects had been placed previously (in the absence of current object information) (13). The fact that these “trace cells” were not involved in the initial experience, but rather reflect the history of experience, might render them particularly well-suited to support offline reactivation of individual learning experiences.

In this fMRI study, we examined patterns of activation across voxels occurring spontaneously during an interval between encoding and retrieval and assessed their similarity to the activation patterns occurring during individual encoding events. The critical question was whether the degree of such reactivation differed for events subsequently remembered relative to those subsequently forgotten, thus directly linking the amount of reactivation to later memory performance. During encoding, participants were presented with random and unique object–scene pairs. During retrieval, a cue–recall test was used that allowed each object–scene trial to be labeled as recalled (R) versus forgotten (F). Importantly, between encoding and retrieval, participants performed a 2-min distracter task during which odd/even number judgments were made (Fig. 1), a task shown to recruit medial temporal lobe (MTL) regions to comparatively little extent despite being cognitively engaging (14). The critical analysis concerned this 2-min delay period, focusing on reactivation in the hippocampus (HIPP) and the entorhinal cortex (ErC), based on the above-mentioned reactivation signals observed in these regions in rodents.

Results

Behavioral Results. The paradigm yielded a balanced distribution of R and F trials [47 and 53%, \((19) = 1.19, P > 0.2\)]. When prompted to describe the paired associate after giving an R response (Methods), descriptions were accurate on 97% (±5%) of the trials.

Significance

How is new information converted into a memory trace? Here, we used functional neuroimaging to assess what happens to representations of new events after we first experience them. We found that a particular part of the medial temporal lobe, a brain region known to be critical for intact memory, spontaneously reactivates these events even when we are engaged in unrelated activities. Indeed, the extent to which such automatic reactivation occurs seems directly related to later memory performance. This finding shows that we can now study the dynamics of memory processes for specific experiences during the “offline” periods that follow the initial learning phase.

Author contributions: B.P.S. and R.N.H. designed research; B.P.S. performed research; A.A. and N.K. contributed new reagents/analytic tools; B.P.S. and A.A. analyzed data; and B.P.S. and R.N.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.P.Y. is a guest editor invited by the Editorial Board.

1To whom correspondence should be addressed. E-mail: bernhard.staresina@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1311989110/-/DCSupplemental.
Does Reactivation Reflect Carryover of Encoding Effects? First, we wanted to ensure that the observed effect in ErC reflects true offline reactivation rather than merely residual activation effects from encoding (akin to “reverberating” or “echoing” effects from the study phase). Such carryover effects could result from two scenarios: (i) R trials might elicit greater blood oxygen level-dependent (BOLD) activation than F trials during encoding, or (ii) R trials might occur systematically later in the encoding run such that delay period activation is more similar to R than F trials due to mere temporal proximity.

To control for scenario (ii), we conducted the same reactivation analysis as described above, but covarying each trial’s BOLD activation level during encoding when deriving the reactivation indices for R and F trials (see Methods for details). Results showed that the reactivation increase for R trials vs. F trials remained significant \( t(19) = 2.47, P = 0.012 \). Moreover, we observed no differences in average BOLD activation for R vs. F trials during encoding \( t(19) = 0.06, P = 0.79, P = 0.438 \). To ensure that we would have sufficient power to detect such differences, we conducted the same BOLD activation comparison for the retrieval data, where we found a strong increase in ErC reactivation for R relative to F trials \( t(19) = 1.42, F(8.2, 154) = 1.26, P = 0.253 \) (Fig. S6).

Turning to our main question of offline reactivation of encoding patterns in our a priori MTL regions of interest [hippocampus (HIPP) and entorhinal cortex (ErC)], we used representational similarity analysis (RSA) (16). In particular, we separately correlated a given region’s across-voxel activation pattern for a given encoding trial with each of the 120 delay period volumes (repetition time = 1 s), thus obtaining a “reactivation index” for each encoding trial throughout the delay period (see Fig. S5 for a schematic overview of the analysis approach). In a first analysis, each participant’s reactivation index for R vs. F trials (averaged across all 120 delay volumes) was entered into a repeated-measures ANOVA, including the factors Region (HIPP, ErC) and Memory (R, F). A significant interaction \( [F(1, 19) = 9.87, P = 0.005] \) reflected the fact that, although there was no evidence for differential reactivation for R vs. F trials in HIPP \( [F(1, 19) = 0.03, P = 0.489] \), there was significant increase in reactivation for R vs. F trials in ErC \( [F(1, 19) = 2.43, P = 0.013] \) (Fig. 2). When including time across the delay period (volumes 1–120) as a factor in an ANOVA on ErC reactivation, there was only a main effect of memory \( [F(1, 19) = 5.92, P = 0.025] \) and no main effect of Time \( [F(8.0, 154) = 1.47, P = 0.170] \), nor a memory \times time interaction \( [F(10.8, 205) = 1.26, P = 0.253] \) (Fig. S6).

fMRI Results: Event-Specific Reactivation During the Delay Period. In general, the object/scene encoding task elicited strong activation in the expected extrastriate regions along the ventral visual stream [e.g., lateral occipital complex (LOC) and parahippocampal place area (PPA)], as well as in the MTL, and “subsequent memory effects” were observed in MTL regions (including the hippocampus), as well as in higher order object- and scene-processing regions, as expected from previous univariate activation studies of subsequent memory effects. These univariate results are fully described in SI Text (Figs. S2–S4) (for a full description of the retrieval data, see ref. 15).

Fig. 1. Experimental paradigm. During encoding, participants were presented with unique object-scene pairs (4 s) and indicated whether each combination was plausible or implausible. During the active delay period immediately following encoding, participants performed a self-paced odd/even number judgment task (2 min). Schematized is the hypothesized differential extent of reactivation for successfully vs. unsuccessfully learned study episodes. During retrieval, participants saw either an object or a scene cue and indicated whether they recalled the corresponding paired associate (6 s).

Fig. 2. Offline reactivation in the entorhinal cortex (ErC). (Left) Hand-drawn anatomical regions of interest (ROIs) shown for one participant. (Right) Reactivation of encoding representations is greater for later recalled (R) than forgotten (F) trials. Bars represent mean ± SE of the condition difference. *P = 0.013, one-tailed paired t test.
differences of 91.8 s vs. 95.8 s relative to the delay period for R vs. F trials, \( t(19) = 1.32, P = 0.202 \). We further quantified the distribution of serial positions (1–32 for any given run) of R and F trials across participants (Fig. S7). Again, there was no systematic bias for R trials to occur more frequently at the beginning or the end of the list, as confirmed by a repeated-measures ANOVA including the factors memory (R, F) and serial position (1–32). There were no effects of memory, serial position, or their interaction (all \( F < 1.6, P > 0.1 \)). The same held true when reducing the factor serial position to two levels by averaging occurrences of R and F trials across the first and last quarter of the run, respectively (trial numbers 1–8 and 25–32), i.e., the segments most reflective of potential primacy and recency effects (all \( F < 2.37, P > 0.1 \)).

Finally, the ErC reactivation increase for R trials again remained significant after covarying out each trial’s temporal distance to the onset of the delay period when deriving the reactivation indices for R and F trials \( \{t(19) = 1.78, P = 0.046 \} \). Also, the same results were obtained when conducting the reactivation analysis separately for the first and second half of the encoding trials: In a repeated-measures ANOVA with the factors memory (R, F) and half (first, second) on ErC reactivation values, there was a main effect of memory \( \{F(1,19) = 5.04, P = 0.037 \} \), but no memory \( \times \) half interaction \( \{F(1,19) = 0.43, P = 0.522 \} \).

**Evidence for Discrete Reactivation Events?** The above analysis showed that the ErC activation patterns throughout the delay period were more similar to the event-specific encoding patterns for R than for F trials. We next set out to assess whether this sustained increase might consist of discrete reactivation events or “bursts” in which individual study trials are reactivated more strongly relative to some baseline measure. To derive such baseline measures of reactivation for each individual trial, we computed a “sham reactivation” index by correlating a given trial’s encoding pattern with the delay period volumes of all other runs. For instance, to derive a sham reactivation index for trial 1 in run 3, we took that trial’s encoding pattern and correlated it with the activation pattern of each volume of each of the other five runs (run 1, 2, 4, 5, and 6 in this example). Given that each delay period consisted of 120 volumes, this gave us a distribution of 600 correlation values under scenarios where no true reactivation would be expected. The average of those sham reactivation values was 0.004 for R trials and 0.005 for F trials, which did not differ significantly \( \{t(19) = 0.53, P = 0.66 \} \). Going back to the actual run from which the current trial was obtained (run 4 in this example), each of the 120 actual reactivation values was then compared with the sham reference distribution, allowing us to convert each reactivation index of a given trial into a Z-value. Discrete reactivation events were then defined as Z-values > 2 (corresponding to a \( P \) value < 0.05), reflecting a local peak in reactivation relative to a trial-specific noise (“sham”) estimate. This procedure revealed potential sequences of discrete reactivation events throughout the delay period for every encoding trial (Fig. 3A). Results showed that there was an increase of reactivation events for R vs. F trials both in (i) the average number of reactivation events per trial \( (11.7 \text{ vs. } 10.6) \) and (ii) the proportion of individual trials exhibiting at least one reactivation event \( (81\% \text{ vs. } 79\%) \) \( \{\text{both } t(19) > 1.85, P < 0.05 \} \) (Fig. 3B). Again, no difference between R and F trials in these metrics was observed in HIPP \( \{t(19) < 0.28, P > 0.38 \} \).

Note that this analysis of discrete reactivation events allows for multiple encoding trials showing reactivation at the same delay period time point. As such, the activation pattern at delay period time point \( t \) across, e.g., 100 ErC voxels may be a product of (i) the same 100 voxels reactivating different encoding events’ representations at the subsecond level and/or (ii) 50 voxels representing encoding event A and 50 voxels representing encoding event B. Although unsolvable at current levels of spatiotemporal resolution of fMRI, both of these scenarios would give rise to multiple encoding events showing a relative reactivation increase in the same BOLD volume.

Lastly, to further corroborate these results, we conducted another complimentary analysis in which reactivation events were defined in an arguably simpler way, namely as correlation values > 1.5 SDs from the current run’s overall mean (i.e., without first deriving sham reactivations across other runs). The same pattern emerged: A significant increase in ErC (and not HIPP) for R vs. F trials in both the number of overall reactivation events \( (8.91 \text{ vs. } 7.89) \) as well as trials showing at least one reactivation event \( (76\% \text{ vs. } 73\%) \) \( \{\text{both } t(19) > 2.04, P < 0.05 \} \).

**Reactivation Beyond the MTL.** Lastly, we conducted an exploratory searchlight analysis (17) to see whether any brain regions beyond our a priori regions of interest would show increased reactivation for R relative to F trials. A sphere (3-voxel radius) was centered on every voxel in each participant’s native space, and reactivation indices were derived as described in the previous section. The resulting reactivation maps were normalized to a standard Montreal Neurological Institute template provided in SPM8. Maps were compared for R trials vs. F trials via a whole-brain paired \( t \) test and thresholded at \( P < 0.001 \), uncorrected (minimum of 10 contiguous voxels). As shown in Fig. 4, only one region emerged from this analysis, located in the left retrosplenial cortex (RsC).

**Discussion**

These results demonstrate that human memory performance correlates with spontaneous reactivation of individual study episodes during an active, postlearning offline period. Although previous studies have observed that brain regions engaged in the study task come back online (6) or increase their functional coupling with the hippocampus (7) during postlearning offline periods, it has remained open whether representations of individual study events reemerge during such periods and whether event memory correlates with this reactivation.
What alternative explanations are there for the reactivation effects we observed? One concern is that R trials might cluster more closely in time to the delay period compared with F trials, such that slow-changing global cognitive, physiological, and/or scanner states are more similar between the delay period and R trials (i.e., induce temporal autocorrelation in the data). Several factors, however, argue against this possibility in our data. First, there was no recency effect for R trials in the encoding data, and the same condition differences were obtained when factoring out each trial’s temporal proximity to the delay period. Also, any unspecified similarity between encoding and the delay period would be expected to gradually decline over time, but there was no differential reactivation effect for R vs. F trials at the beginning vs. the end of the delay period (Fig. S6). Moreover, any spurious correlation unrelated to actual reactivation would be expected to show up across the entire brain, unlike the highly region-specific effects observed here in ErC and the retrosplenial cortex (even at reduced thresholds) (Fig. S8).

Another possibility is that the increased pattern reactivation for R trials may in fact reflect commonalities between the delay period and R trials other than event-specific representations. For instance, one could imagine that R trials (but not F trials) are associated with a certain attentional state during encoding and that the active delay period puts participants in the same attentional state again. However, this explanation, and related explanations, are controlled for by the sham reactivation analysis (Fig. S5B). To reiterate, in the sham reactivation analysis, we took a given trial’s encoding pattern and correlated it with the delay period volumes of all other runs. Thus, if any type of similarity between R trials and general features of the delay period task were to drive our reactivation indices, those similarities should be preserved across runs, and thus the sham reactivation for R trials should also be greater than that for F trials. As reported in Results, this was not the case. In fact, there was a significant interaction between memory (R, F) and reactivation (real, sham), reflecting the fact that the difference in reactivation indices for R vs. F trials were significantly greater in the real compared with the sham analysis \( F(1,19) = 5.94, P = 0.025 \).

Another (unwanted) factor that may have driven similarity between encoding and the delay period is the use of the same odd/even numbers task for encoding baseline trials and for the delay period task. For instance, if a given encoding trial is followed by a baseline trial that includes the number sequence “8-97-42,” reemergence of that same number sequence (or part thereof) during the delay period may have led to implicit reactivation of the adjacent encoding trial, akin to the finding that sub consciousness presentations of sounds associated with study items during offline periods bolsters later memory for those items (8). However, the above-mentioned sham reactivation analysis would again control for this possibility—if, e.g., trial 1 in run 1 is followed by a long baseline trial, then the likelihood that some of those numbers will reoccur in the delay period of other runs is also increased. Moreover, as detailed in SI Text, there was no evidence that the amount of individual number overlap between a given run’s encoding and delay phases was related to memory performance.

Key to our interpretation of the results as reflecting endogenous reactivation is the assumption that participants did not actively rehearse the study material during the delay period. Although the theoretical distinction between replay and covert retrieval is certainly a subtle one, our control analyses argue against such intentional rehearsal strategies: Not only did delay period performance (both in terms of RTs and accuracy) and ErC reactivation remain constant across runs (indicating that participants did not adjust their performance to better accommodate active rehearsal), but none of the participants tested in a follow-up behavioral study claimed that they rehearsed the study material (SI Text).

Finally, how meaningful are correlation values of \( \sim -0.04 \) (Fig. 2)? First, it is of course difficult to directly interpret correlation values in (noisy) fMRI data (consider that the typical evoked BOLD response comprises \( \sim 1\% \) signal change from baseline). More importantly, the reactivation values we obtained here are in good agreement with correlation values reported in previous studies that compared MTL activation patterns between encoding and retrieval [e.g., \(-0.06 \) during successful recall in ref. 18 or \(-0.04 \) in ref. 19, or that compared MTL activation patterns across different stimulus exemplars during encoding (\(-0.06 \) in ref. 20)].

Another issue concerns the extent to which the offline reactivation effects we observed are indeed relevant for later memory performance. That is, how much does this effect contribute beyond the subsequent memory effects (SME) typically observed at encoding (21–25) (see SI Text, Univariate Analysis of Encoding and Delay Period Data for subsequent memory effects in the current paradigm)? For one, both SME and offline reactivation are merely correlational, and theoretically there is no reason to assume that one is more causal for later memory performance than the other. And just as one might argue that offline reactivation reflects an epiphenomenal aftermath of successful encoding, it may be conceivable that successful encoding merely “tags” which study events will be reactivated in the subsequent rest period, and it is that reactivation that transduces a working memory representation into long term memory. In any case, experimentally (and covertly) reactivating individual study events during an active delay period was recently shown to bolster recall for those events (8), suggesting that offline reactivation is indeed of functional significance. That said, we note that 2 min is a comparatively short delay and that further work is needed to extend our conclusions to longer delays between study and test. There is, however, little evidence that, once working memory/rehearsal capacities are exceeded, memory signals would vary as a function of the delay length. For instance it has been shown that, within the MTL, subsequent memory effects are invariant to whether memory is tested 30 min or 48 h later (24).

At first sight, it may seem surprising that we did not observe reactivation effects in the hippocampus, given the eminent replay phenomena reported in this region in rodent models (10, 25). However, the transient nature and regional specificity of replay within the hippocampus [replay tends to occur in a time-compressed fashion (25) and is typically observed in the CA1 or CA3 subfields (26)] may impede detection of such events with fMRI. Also, in human intracranial recordings, ripple events (which have been intimately linked to replay) (see ref. 27 for review) have been observed both in HIPP as well as in ErC (28), but interestingly the number of ripple events during a postlearning nap correlated with later memory performance only for ErC ripples and not for HIPP ripples (29). It thus remains an open question whether we did not observe HIPP reactivation effects because of the spatiotemporal resolution of the conventional fMRI used here is too low to detect such events in HIPP, or because memory-related offline reactivation is more robustly manifest in ErC in humans. Finally, we might have had reduced power to detect HIPP effects due do the relatively short delay between study and test—although note that we did observe subsequent
memory effects in HIPP during encoding (supplemental material) and recall success effects during retrieval (15).

What may put ErC in a privileged role to reveal reactivation effects in our paradigm? Beyond the pivotal role of ErC grid cells for spatial navigation (30–32), both object cells that code for individual object stimuli at the exploration phase and trace cells that code for an object’s past location have been described in this region (13). Of course, a direct mapping of those signals to the human brain is difficult, not least because of methodological considerations. That said, at the resolution of fMRI voxels, it is likely that the signal we observe—to the extent that neural firing is expressed in BOLD signal—would be a mixture of contributions from both of these cell types. Although it has not been shown whether trace cells have corresponding object cells (i.e., for every object cell there is a linked trace cell), such a pairing, especially if among adjacent cells, would express itself in increased correlations across encoding and the delay period within the same voxel, as observed here.

Furthermore, ErC cells in nonhuman primates were found to maintain information during a delay period for both objects and places (33). Although those cells were identified in a short-term memory paradigm, they may well serve similar functions during longer delays and in the absence of active rehearsal (ST Test and TST) in humans. In that sense, our stimulus material, consisting of object–scene pairs, might be particularly well-suited to elicit ErC engagement. That is, object and scene information converge on ErC (34–36), such that this region is in a privileged position to reactivate the entire object–scene study event (37). Future studies using different stimulus types will shed light on the potential domain-generality of ErC reactivation effects. Also, it will be interesting to test whether the reactivation effects we observed apply specifically to associative memory, as assessed here, or whether nonassociative expressions of memory benefit from object reactivation to similar extents.

When extending our reactivation analysis beyond the MTL, the only other region in which we observed differential reactivation for R vs. F trials (although note that we had limited brain coverage) (Fig. S8C) was the retrosplenial cortex (RsC). The only study to date that performed simultaneous fMRI while externally reactivating study material (albeit conducted during slow wave sleep and reactivating a global study context rather than individual memory-related processes (39))—numbers of memory-related processes may already be observable during consolidation (41, 42), it remains to be determined whether early consistent with the purported role of this region in systems consolidation (43, 44). The only study so far to use object reactivation to explore the potential domain-generality of ErC reactivation effects. Also, it will be interesting to test whether the reactivation effects we observed apply specifically to associative memory, as assessed here, or whether nonassociative expressions of memory benefit from object reactivation to similar extents.

When extending our reactivation analysis beyond the MTL, the only other region in which we observed differential reactivation for R vs. F trials (although note that we had limited brain coverage) (Fig. S8C) was the retrosplenial cortex (RsC). The only study to date that performed simultaneous fMRI while externally reactivating study material (albeit conducted during slow wave sleep and reactivating a global study context rather than individual memory-related processes (39))—numbers of memory-related processes may already be observable during consolidation (41, 42), it remains to be determined whether early consistent with the purported role of this region in systems consolidation (43, 44). The only study so far to use object reactivation to explore the potential domain-generality of ErC reactivation effects. Also, it will be interesting to test whether the reactivation effects we observed apply specifically to associative memory, as assessed here, or whether nonassociative expressions of memory benefit from object reactivation to similar extents.

Methods

Participants. Twenty (11 female) right-handed native English speakers with normal or corrected-to-normal vision participated in the experiment (mean age, 25 y, range, 22–32). Informed consent was obtained in a manner approved by the Cambridge Psychological Research Committee, and participants were paid for their participation.

Experimental Design. The stimulus material consisted of 384 color pictures (44), half of which (192) depicted objects and half of which depicted scenes (16 additional pictures were used for practice). The experiment consisted of six runs, each run containing three blocks: an encoding block, a delay period, and a retrieval block (Fig. 1). Some of the trials performed during object reactivation were performed continuously across the three blocks, with short unscanned breaks between runs. The current analysis focused on the encoding and delay period data. During each encoding block, participants were presented with 32 unique object–scene pairs. The pairing of objects and scenes was randomized across participants. Object and scene pictures were each presented in a 250 × 250 pixels frame placed to the left and right of the screen center. During the trials, the trials were presented on the left of the screen center and the scene to the right, with the reverse order during the other half of the trials. The trial duration was 4 s and, for the last 0.5 s, the picture pair was replaced with a fixation cross (responses were still recorded), alerting participants that another trial would appear shortly. The encoding task was to indicate via button press whether the given object–scene pair is plausible or implausible, i.e., likely to appear in real life or nature (45). “Plausible” responses were given with the index finger and “implausible” responses with the middle finger. Across participants, use of left vs. right hand was counterbalanced (but the finger assignment was held constant). Object–scene encoding trials were interleaved with an active baseline task (14). Here, random numbers between 0 and 100 were shown, and participants pressed the index finger key for even numbers and the middle finger key for odd numbers. As soon as a response was given, another random number was shown. The response time for each number was self-paced, and participants were encouraged to perform this task as fast as possible without sacrificing accuracy. Each encoding block lasted about 3 min.

After the last encoding trial, participants saw a transition screen for 16 s (including a progress bar), alerting them to the upcoming delay period. The transition was inserted to allow for the return of the hemodynamic response to baseline levels. During the delay period, participants again performed the odd/even task above for 2 min. Performance accuracy was reported to participants on the computer screen following the completion of the task to encourage accuracy.

At the end of the delay period, another 16-s transition screen alerted participants to the upcoming retrieval block. Each retrieval block consisted of 32 trials, each trial lasting 6 s. For a given trial, participants saw either the object or the scene of a given object–scene pair from the previous encoding block and were asked to indicate whether they remembered the corresponding paired associate (“recall”, index finger) or not (“forgot”, middle finger). Half of the cues (16) were object pictures, the other half scene pictures. To ensure that participants gave recall responses when they indeed recalled the correct paired associate, we asked them to verbally describe the target after ~10% of the recall responses (for more details on the retrieval phase, see ref. 15). The retrieval block lasted ~6 min.

fMRI Analysis. Data were analyzed using SPM8 (www.fil.ion.ucl.ac.uk/spm/). During preprocessing, images were corrected for differences in slice acquisition timing, followed by motion correction across all runs. All events of a given run were modeled in a mass-univariate general linear model (along with nuisance regressors for head motion, low frequency scanner drift, and mean signal per run). Each encoding and retrieval trial was modeled as an individual impulse regressor convolved with a canonical hemodynamic response function. The other events in a given run, including (i) the transition phase between encoding and delay, (ii) the transition phase between delay and retrieval, and (iii) the delay period itself, were modeled using unconvolved stick functions for each volume. That is, for the critical delay period, there are no a priori (experimentally controlled) times at which to expect the onset of a reactivation response. Thus, the tacit assumption is that any acquired delay period volume may reflect the peak response of an HRF, triggered by a neural reactivation event. The resulting beta values for all events were transformed into t values. The transformation from beta- to t-values was done by dividing each voxel’s beta parameter estimate by the SE of that voxel’s residual error term after fitting the first-level general linear model.
The benefit of this transformation for correlational analyses is in down-weighting noisy voxels and increasing the sensitivity of pattern-information analyses (16; 47). Note, however, that results were not contingent on this transformation (SI Text, Reactivation Control Analyses). No smoothing or normalization was performed on the echo planar imaging data that entered the GLM.

Delay period reactivation of encoding events was quantified via representational similarity analysis (RSA) (16) (Fig. S5A). Before entering t tests or ANOVAs, reactivation indices were Fisher z transformed. For ANOVA factors with more than one numerator degree of freedom (df), we used a Greenhouse–Geisser correction for non-sphericity of the error. Measures for R and F trials were combined across the six runs via an average weighted by the number of R and F trials, respectively, in each run.

We analyzed the data in HIP and Erc using hand-drawn, participant-specific ROIs, based on the individual structural image (Fig. 2). Anatomical demarcation was done according to Insausti et al. (48) and Pruessner et al. (49). Left and right hemisphere ROIs were combined for analyses. Note that no spatial smoothing was performed on the data, ensuring that there was minimal signal overlap between the regions. The average number of voxels in the anatomical masks was 179 (range 137–253) for HIP and 101 (range 73–153) for Erc. Given our clear a priori hypothesis that memory would benefit from increased offline reactivation, direct comparisons between R and F trials were conducted via one-tailed, paired-samples t tests for our regions of interest.

To covary out certain factors of no interest at the individual participant and run level, we devised a general linear model in which the dependent variable was the 32 reactivation indices for the encoding trials. The predictors were separate regressors for (i) R trials (a 32 × 1 vector coding 1s for R trials and 0s for F trials), (ii) F trials (a 32 × 1 vector coding 0s for R trials and 1s for F trials), (iii) the (mean-centered) covariate to be controlled for (e.g., the serial position of each trial), and (iv) a constant term. The parameter estimates for the first two regressors then reflect the average contribution of R and F trials to the observed reactivation indices, respectively, after accounting for other possible factors, such as the serial position, trial-specific BOLD activation, or plausibility judgments during encoding.

ACKNOWLEDGMENTS. We thank Elisa Cooper for help with data acquisition and Anthony Wagner for his support. This work was supported by a Sir Henry Wellcome Postdoctoral Fellowship (to B.P.S.), the United Kingdom Medical Research Council Programs MC-A060-SPR10 (to R.N.H.) and MC-A060-SPR20 (to N.K.), a European Research Council Starting Grant (to N.K.), and Rubicon Grant from the Netherlands Organization for Scientific Research (NWO) 825.10.023 (to A.A.).
Supporting Information

Staresina et al. 10.1073/pnas.1311989110

SI Text

Delay Period Performance Across Runs

First, in a repeated-measures ANOVA on delay period RTs, there was no evidence of a main effect of the factor Run \([F(2.36,44.78) = 1.26, P = 0.30]\), even when specifically testing for a linear trend \([F(1,19) = 1.30, P = 0.268]\) or when directly comparing reaction times (RTs) for the first half of the experiment with the second half \([\text{first half average, } 1.30 \text{ s; second half average, } 1.27 \text{ s}; t(19) = 1.13, P = 0.273]\) (see Fig. S1 for a distribution across runs and participants). Second, there was no correlation between a given run’s delay period performance (odd/even RT) and that run’s recall performance when pooling across runs and participants. (Spearman’s \(r = -0.02, P = 0.789\)). Likewise, odd/even judgment accuracy remained constant across runs, with an overall average of 97% (±2%) and again there was no evidence in a repeated-measures ANOVA of a main effect of the factor run \([F(3.25,61.62) = 2.05, P = 0.11]\), even when specifically testing for a linear trend \([F(1,19) = 0.81, P = 0.379]\) or when directly comparing accuracy for the first half of the experiment with the second half \([\text{first half average, } 96.43\%; \text{second half average, } 96.83\%; t(19) = 0.89, P = 0.384]\). Furthermore, there was no correlation between a given run’s delay period accuracy and that run’s recall performance when pooling across runs and participants (Spearman’s \(r = -0.04, P = 0.701\)).

Univariate Analysis of Encoding and Delay Period Data

Figs. S2 and S3 show results from standard univariate task- and subsequent memory analyses. For data from the retrieval phase, see Staresina et al. (1).

The odd/even task engaged comparable sets of regions during encoding and the delay period. This result is shown in Fig. S4, in which Fig. S4A shows the contrast “encoding task < encoding BL” (baseline), Fig. S4B shows the contrast “encoding task < delay”, and Fig. S4C shows their overlap. No regions survived when exclusively masking one contrast with the other. In short, there was no evidence for the odd/even task recruiting different areas as a function of the experimental phase.

Reactivation Control Analyses

Did Reactivation Already Occur During the 16-s Transition Period Immediately Following the Encoding Part? The same reactivation analysis as performed for the delay period showed no difference in entorhinal cortex (ErC) reactivation for recalled (R) vs. forgotten (F) trials \([t(19) = 1.34, P = 0.099, \text{one-tailed}]\) during the transition period, consistent with participants’ report in the questionnaire of the additional behavioral study (see Behavioral Results in the main text).

Is Reactivation Driven by the Similarity of the Delay Period Task with the Encoding Baseline Condition? One possibility is that parameter estimates of encoding trials are at least partially affected by the interspersed odd/even baseline trials. We thus conducted the same reactivation analysis after adding regressors to the first-level models that captured the duration of each baseline trial. This did not affect the reactivation effect in ErC \([\text{greater reactivation for R vs. F trials}; t(19) = 2.11, P = 0.024]\).

Does the Plausibility Judgment During Encoding Affect Reactivation Above and Beyond Subsequent Memory? The overall ratio of plausible to implausible responses was 29 and 71%; i.e., most object/scene combinations were judged implausible by the participants. This relatively low number of plausible trials precluded separate memory analyses for those two response types because many participants had <10 plausible F trials. Importantly, however, in the reactivation analysis, we controlled for plausibility judgments by adding a regressor coding for judgment type (plausible or implausible) in a participant- and run-specific general linear model in which we derived the reactivation indices for R and F trials, and the effect of increased reactivation for R vs. F trials in ErC remained significant \([t(19) = 2.27, P = 0.017]\). Moreover, there was no difference in reactivation between remembered plausible trials and remembered implausible trials: \(t(19) = 0.91, P = 0.374\).

Use of t Values for Representational Similarity Analysis and Controlling for Outliers. As mentioned in Methods, individual trial parameter estimates were converted to t values before calculating pattern correlations. Although this method has become “best practice” in representational similarity analysis (RSA) approaches (2, 3), it should be noted that the same results in ErC were observed when using beta estimates instead of t values [reactivation for R vs. F trials: \(t(19) = 2.41, P = 0.013\). Although this transformation to t values down-weights noisy voxels and thus reduces the likelihood of outliers in the data, we additionally conducted our reactivation analysis after removing voxels whose t values exceeded ±2.5 SD of the mean across all trials and voxels within a run. The same reactivation effect was observed in ErC: R > F trials in ErC, \(t(19) = 2.09, P = 0.025\), as was also the case when using an outlier criterion of ±5 SD of the mean, \(t(19) = 2.46, P = 0.012\).

Is the Reactivation Effect Stable Across the Six Experimental Runs? Reactivation indices for R and F trials in ErC did not differ across runs: First, in a repeated-measures ANOVA with the factors memory (R, F) and run (1–6), there was only a main effect of memory \([F(1,19) = 5.49, P = 0.030]\) and no effect of run \([F(4.18,79.34) = 0.67, P = 0.620]\) and no memory x run interaction \([F(4.32,82.16) = 0.53, P = 0.726]\). Second, the same results were obtained when averaging across runs 1–3 (first half) and runs 4–6 (second half), respectively: a main effect of memory \([F(1,19) = 5.37, P = 0.032]\) and no effect of half \([F(1,19) = 0.25, P = 0.63]\) and no memory x half interaction \([F(1,19) = 0.85, P = 0.37]\).

Behavioral Study

We conducted an additional behavioral study in which the experimental parameters were identical to the fMRI version, with the exception that we only used two runs and included a questionnaire after completion of the second run. In this questionnaire, we explicitly asked participants whether they mentally rehearsed the study material during the delay period (and/or any of the transition periods) (Table S1).

Results closely replicated data from the functional MRI (fMRI) study. There was a balanced distribution of R to F trials \([52–48%, t(9) = 0.45, P = 0.66]\), and 90% (±14%) of catch trials were given accurate responses. Importantly, like the fMRI study, there was no evidence that participants adjusted their delay period performance across the runs. The average RT per digit was 1.28 s during run 1 and 1.27 s during run 2 \([t(9) = 0.20, P = 0.849]\), and the average odd/even accuracy was 93.7% during run 1 and 94.4% during run 2 \([t(9) = 0.89, P = 0.398]\).

Again, no serial position effects were observed for R vs. F trials. First, the mean positions of R and F trial in the encoding phase were comparable \([16.8 \text{ for } R \text{ and } 16.4 \text{ for } F; t(9) = 0.50, P = 0.63]\). Second, there were no differences in the serial distribution of

Staresina et al. www.pnas.org/cgi/content/short/1311989110
events across participants. In a repeated-measures ANOVA with
the factors emory (R, F) and serial position (1–32), there was no
effect of memory, serial position, or their interaction (all \( P >
0.19 \)).

Critically, as shown in Table S1 below, participants’ self-report
confirmed what the behavioral data indicated: no evidence that
active rehearsal of the study material took place during the delay
period or the transition period between study and delay.

We also tested the possibility that—given that the odd/even task
was used both during encoding baseline trials and during the delay
period—the numbers presented during the delay period may serve as
incidental retrieval cues for the study material (akin to findings that
experimentally cueing memoranda during sleep bolsters memory
performance) (4, 5). To this end, we recorded each individual
number occurring during both the encoding phase and the ensuing
delay phase and quantified their overlap. There was, however, no
difference between the extent of number overlap for the run with
better memory performance compared with the run with worse
memory performance [average number overlap for worse run, 61.6;
average number overlap for better run, 64.2; \( t(9) = 0.99, P = 0.35 \)].

1. Staresina BP, Cooper E, Henson RN (2013) Reversible information flow across
the medial temporal lobe: The hippocampus links cortical modules during memory

connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.

classifiers and response normalizations for pattern-information fMRI. Neuroimage
53(1):103-118.


by reactivating them during sleep. Science 326(5956):1079-1079.

Fig. S1. Distribution of reaction times (RTs) during the odd/even delay task across runs. Each color bar represents a particular participant. Connected diamonds reflect the mean across participants, illustrating that RTs did not increase across runs (as would be expected if participants adopted a rehearsal strategy).
**Fig. S2.** Task-related brain activation during encoding. The maps show regions emerging from the contrast task > baseline, thresholded at a minimum of five contiguous voxels exceeding an uncorrected P value of 0.001. Upper row shows global intensity projections; Lower row shows clusters overlaid on the mean anatomical image across participants (Montreal Neurological Institute coordinates of the three displays: $x = 27, y = -49, z = -11$).

**Fig. S3.** Subsequent memory effects. The maps show regions emerging from the contrast subsequent R responses > subsequent F responses, thresholded at a minimum of five contiguous voxels exceeding an uncorrected P value of 0.001. Upper row shows global intensity projections; Lower row shows clusters overlaid on the mean anatomical image across participants (MNI coordinates of the three displays: $x = -21, y = -16, z = -23$).
Fig. S4. Brain regions activated by the odd/even task relative to the object/scene task during encoding (A) and the delay period (B). C highlights the overlap of A and B, illustrating that the odd/even task recruited the same areas irrespective of the experimental phase (MNI coordinates of the three displays: $x = 48, y = 2, z = -32$). No regions survived when exclusively masking one contrast with the other. Note that the entorhinal cortex (ErC) was similarly engaged during encoding, baseline, and delay task: In a repeated-measures ANOVA, there was no effect of the factor condition (encoding task, encoding baseline, delay) on ErC activation levels [$F(1.07,20.24) = 0.20, P = 0.672$].
Fig. S5. Schematic of the reactivation analysis. (A) The matrix depicts the correlation of all 32 encoding trials in a given run, grouped into later recalled (R) and later forgotten (F) trials, with each volume of the subsequent delay period (120 volumes). The inset highlights the correlation values of interest, and grayscale pixels schematize a greater average reactivation index (across encoding trials and delay volumes) for R vs. F trials. (B) Sham reactivation analysis. For a given encoding trial (here schematized for a trial from run 3), reactivation indices were calculated using all other runs’ delay periods.

Fig. S6. Reactivation time course across the delay period, averaged across all participants and runs, for later recalled trials (R, green) and later forgotten trials (F, red). Shaded areas show SEM across participants. Note that the same results were observed when excluding the first 20 volumes from the analysis: a main effect of memory \(F(1,19) = 5.88, P = 0.025\) and no main effect of time \(F(10.46,198.83) = 1.30, P = 0.230\), nor memory x time interaction \(F(12.78,242.76) = 1.41, P = 0.155\).
Fig. S7. Distribution of R (green) and F (red) trials along serial positions at encoding across participants. There was no bias for R trials to systematically occur during the beginning or the end of the study list (see Results for statistical assessment).

Fig. S8. Searchlight analysis results shown at a reduced threshold, $P < 0.005$ uncorrected. (A) Maximum intensity projection of the resulting clusters, highlighting left entorhinal cortex (solid circle) and left retrosplenial cortex (dashed circle). (B) Clusters are superimposed on the average anatomical image across participants (MNI slice coordinates: $x = -9$, $y = -7$). (C) Brain coverage (white) for the searchlight analysis, shown for slices ranging from $x = -70$ to $x = 0$. 
Table S1. Questionnaire

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did you actively rehearse the object/scene pairs during the delay portion (the 2-min odd/even task between encoding and retrieval)?</td>
<td>0</td>
</tr>
<tr>
<td>Did you actively rehearse the object/scene pairs during the first transition period (between encoding and the 2-min odd/even task)?</td>
<td>0</td>
</tr>
<tr>
<td>Did you actively rehearse the object/scene pairs during the second transition period (between the 2-min odd/even task and retrieval)?</td>
<td>1</td>
</tr>
<tr>
<td>When you gave a “remember” response during retrieval, do you think it is more likely that you “mentally practiced” those pairs during the delay or transition periods?</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig. S9. Encoding baseline schedule. Distributions of (Upper Left) serial position of task trials, (Upper Right) serial position of baseline trials, (Lower Left) number of baseline trials (there were always 32 task trials), and (Lower Right) duration of baseline trials.